Gesture Interaction with Spatial Audio Displays: Effects of Target Size and Inter-Target Separation

Georgios Marentakis, Stephen A. Brewster

Glasgow Interactive Systems Group, Department of Computing Science,
University of Glasgow, Glasgow, G12 8QQ, UK
{georgios, stephen}@dcs.gla.ac.uk www.audioclouds.org

ABSTRACT

This paper presents the results of an experiment comparing two spatial audio display segmentation techniques by investigating the relative salience of target width versus distance to target in a gesture based spatial audio selection task. The first technique, MINIMAL, occupies as little of the display area as possible with sounds placed as close to each other as possible. The second technique, MAXIMAL, occupies all the available display area and sounds are placed as far apart as possible and the associated display area assigned to each sound is allowed to grow. Ratios of distance to target to target width were kept constant in both displays to investigate the relative salience of distance to target versus target width in the sound selection task. Participants performed an orientation based pointing task to select an audio display element in the presence of distracting sounds. Results show that the maximal strategy results in faster and more accurate interaction. Target width was found to have significantly more impact on time ratings than distance to target. Time and accuracy ratings indicate that deictic gesture interaction with a spatial audio display is a robust and efficient interaction technique.

1. INTRODUCTION

Spatial audio displays are important from a usability point of view because they enable eyes free interaction. This is due to the fact that audio displays work primarily through our hearing sense. Thus display presentation does not require a computer screen and for this reason such displays are highly portable. Therefore audio displays are suitable in areas such as wearable and mobile computing. This is also because audio displays are well suited for the purpose of monitoring tasks when visual contact is not possible or not convenient due to lack of screen space. Finally, spatial audio displays can enable human computer interaction for the visually impaired. Despite the possible application domains, the design of spatial audio displays has not been examined in detail.

By the term audio display, we refer to an interface, in which display elements are presented using audio. When the display elements are allocated in different positions in space, it is customary to refer to the display as a spatial audio display.

Spatial audio technology is often used extensively in spatial audio displays. This technology enables people to perceive a sound as emitting from a certain direction in space by applying certain signal transformations to the sound signal. One way of accomplishing this is by filtering through Head Related Transfer Functions (HRTFs). HRTFs are measured empirically and capture the properties of the path to the inner ear, including the effects of the outer ear. When applied to a signal HRTFs result in the signal being perceived as emitting from a given direction in space. HRTF filtering can be implemented in real time and can thus provide a portable way to produce spatial audio.

Spatial audio is useful in the design of audio displays since it can indicate to a user the position/direction of a display element. It has also been found that it increases the intelligibility of concurrent audio streams, due to the phenomenon of the Cocktail Party Effect [2]. For these reasons, 3D sound facilitates the display element allocation in space which is a quite powerful tool extensively used in direct manipulation visual interfaces. The idea of utilizing spatial alignment of elements in audio displays has led to the concept of audio windows, as introduced by Cohen and Ludwig in [15]. Applications of the audio windows concept can be found in Brewster *et. al.* [6] and Savidis *et. al* [21],where, spatial alignment of audio display elements has been used to enable interaction with audio menus. These systems use either radial pie menus or grids to present the elements of the display and the user interacts with the system using gestures such as pointing or nodding.

Other audio displays designs take advantage of linguistic information and are based on speech synthesis and recognition. Such displays have been used to present textual information as treated in Raman [19]. Spatially positioned display elements have also been used in such displays to facilitate document browsing and indexing. Such designs have been reported in Goose [12] and Kobayashi [13]. The former system deals with the presentation and navigation in textual documents. The latter deals with the presentation of web page content and allows link traversal. Space based display element allocation has been used in these displays to metaphorically refer to certain structural elements in the document and also to sonify intra and inter document link traversal.

Users can interact with spatial audio displays in a number of ways. In Cohen [9] gestures were used to control the spatial audio system. Users could point to sounds to select them, catch sounds and drag them along the interface and throw sounds to different display locations to accomplish interaction. Other gesture control mechanisms include nodding. Nodding has been used for navigating in a virtual audio world as in Schmandt [23] and for accomplishing display element selection in a pie menu around a user's head. as in Brewster *et al.* [6]. Other interaction techniques include using a virtual pointer controlled by a mouse device, touch tablets as in Kobayashi [13] or using speech recognition as in Schmandt and colleagues [22, 24]. Audio display content so far has been based on Auditory Icons [11], Earcons [4, 7] and speech.

This research forms part of the AudioClouds project (www.audioclouds.org). The project is focused on Three-Dimensional Auditory and Gestural Interfaces for Mobile and Wearable Computers. The project examines spatial audio displays and aims at assessing their applicability. It also aims at enhancing the understanding of the design of such displays and

quantifying the effects of different design alternatives. Audioclouds are also investigating gesture based control that can be used in these types of interfaces. Audio feedback has been examined as a means to enhance gesture learnability and repeatability and help people cope with rich gesture vocabularies.

The question being investigated in this paper is how to allocate display area to display elements in a uniform manner. Such a decision directly affects target size and target separation and therefore interaction. The problem is tackled from a theoretical point of view by studying the properties of the presentation modality and of the interaction technique, the display is equipped with. In the presented study, spatial audio is used for display and physical pointing to a spatial sound source is used as the interaction technique. Because of the fact that spatial audio indicates display element position the resulting interaction is similar to the interaction observed in visual direct manipulation systems displays. Aspects of psychoacoustics and interaction design are therefore relevant to the experimental question. The rest of the paper presents relevant aspects of the relevant fields of the study and the design and implementation of an experiment that was designed to give answer to the problem of audio display element allocation.

2. PSYCHOACOUSTICAL ASPECTS OF DISPLAY DESIGN

Two of the most important display design choices are target size and target separation. Using suitable values the above choices can help overcome deficiencies of the interaction. A prominent factor that can be accounted for by target width is the inherent variability of the action of pointing to a directional audio cue. When interacting in a spatial audio display, users have to rely on the localization capability of our auditory system, which even with real world sounds, is limited [5]. The uncertainty inherent in estimating the location of sound events is termed Localization Blur. Localization Blur depends on the position of the sound source, its spectral content and the temporal variation of the spectral content of a sound source. It ranges from ±3.6° in the frontal direction, to ±10° on the left/right directions and ±5.5° to the back of a listener under well controlled conditions and real sound sources presented by loudspeakers [5]. In current audio environments, localization blur is far higher, in the order of 20° to 30° as has been reported by Wenzel [26, 27], both when using individualized and when using representative HRTF functions. The particular localization blur measurements have been calculated as a gross average on localization judgments that include target sound positions varying both in elevation as well as in azimuth. For targets constrained in elevation on the horizontal plane, localization blur falls below 20° degrees and can be as low as 10° depending on the direction of the sound event and individual differences [3, 8]. Both in the real world as well as in the virtual audio display case localization blur is higher on the sides of a listener. In effect, sounds emitting from diagonal directions tend to be perceived with a side bias.

Along with the localization error another phenomenon that often occurs is confusions with respect to the direction of the sound event. Front-Back and Up-Down confusions have been often observed during evaluation of virtual spatial audio systems. Front-back confusions happen when sound events defined to appear in the front are perceived as appearing to the back and vice-versa. Confusions with respect to sound source elevation are also common, in which case sounds programmed to appear above the horizontal plane are perceived to appear below and vice-versa. This phenomenon is attributed to the so called cone of confusion (the area inside which inter-aural time and intensities differences are not associated uniquely with a single sound position).

A number of ways have been investigated by researchers to improve localization performance. A prominent example is ‘active listening;, the process of updating the sound scene in real-time based on information with respect to the user’s head orientation relative to the sound sources. This particular technique has been found useful in significantly reducing confusions, however they are not eliminated. It is also the case that the implementation of head tracking in real-time is computationally intensive and not always available in all application domains.

Another potential problem is that most 3D audio rendering systems work by using representative HRTF’s. This is believed to be hindering localization performance. There is some debate on whether individualized HRTF’s assist in disambiguating the direction of sound events with respect to confusions. Wenzel [27] found that using individualized HRTF’s significantly reduced the number of confusions as opposed to Bronkhorst [8] where no such effect has been observed.

Reverberation has also been used as a tool to improve performance in spatial audio displays. Begault *et al.* {Begault R., 2001 #18} found that reverberation proved to be beneficiary for reducing azimuth errors, an improvement of about 6° on average; however reverberation significantly degraded localization performance with respect to elevation. In addition, reverberation was found to assist listeners in overcoming the problem of intracranial localization that is the phenomenon where a sound stimulus is perceived as emitting from within or at the edge of a heads. This is a phenomenon that occurs mainly when sounds are presented to a listener using headphones.

The presence of high frequency components in the sound stimuli was also found to be beneficiary to localization. In [8], it has been reported that the presence of high frequency components in the stimuli can improve localization judgments. Localization error was reduced from 20° on average for a cut-off frequency of 4 kHz to 10° on average for a cut-off frequency of 16 kHz.

Despite the benefit from improvements in the spatial audio rendering technique localization error remains quite high. Due to its magnitude, it is therefore expected that a mismatch between the perceived and actual sound positions will occur. A side effect of the user uncertainty in virtual audio environments is increased homing times. This phenomenon is mainly associated with egocentric spatial audio displays that make use of active listening. This has been observed in a study by Loomis *et al.* [14], where participants were asked to locate a sound by physically moving to it. The sound signal was updated in real time using distance and orientation cues depending on the user relative position with respect to the target. The authors found that people could quickly get to the target sound source however; there was a significant delay until participants were convinced that they were actually on target.

The review points to the fact that sound localization in spatial audio displays is not particularly accurate. This will cause a great deal of ambiguity with respect to display element positions, a fact that will make interaction problematic. Users of deictic spatial audio displays will eventually feel uncertain with respect to the result of their actions and user frustration will become a problem in addition to increased reaction times and error rates. A way to compensate and overcome localization ambiguities is through audio feedback. Using audio feedback to mark sound positions in the display will help to disambiguate

sound positions and overcome the deficiencies that are associated with virtual audio environments.

The usability of audio feedback for on-target confirmation is not restricted to audio selection tasks. In [1], Akamatsu *et. al.*, compared four different feedback types for people selecting a visual target using a mouse-type device. The authors presented an experimental comparison between audio, tactile, normal (i.e. no feedback), color and combined feedback, which was presented to the participants once they were inside the target area. Results showed no differences for the time required to reach the target, however, the type of feedback significantly affected the final positioning times, i.e. time elapsed from the cursor entering the target to selecting the target). The ranking was tactile, combined, audio, color and normal, however it should be noted that the differences were in the order of less than 50 ms. Although this study was performed in a visual display it points to the fact that audio feedback can be successful in accounting for increased homing times problems. It is worth mentioning that audio feedback is also necessary from a usability point of view in order to indicate the display elements states, for example whether an element has focus or is selected. The choice of augmenting the audio display with audio feedback is therefore a natural one.

According to the review, directional cues only are not sufficient to enable effective and efficient selection of a spatial audio sound source by means of a pointing gesture. We would expect significant error rates due to localization ambiguity and increased final positioning times. From a design point of view these problems directly affect interaction with a spatial audio source, both in terms of display scalability and in terms of reaction times in interaction with the display and have to be tackled by design. Feedback can serve as a design option that can be helpful in overcoming the particular problems.

To conclude, the literature review shows that problems with sound localization in spatial audio displays require the use of feedback in order to overcome the time and accuracy problems. We therefore use feedback in the spatial audio display of this experiment to overcome potential interaction problems.

To proceed with the design of the display and to answer our experimental question we need to decide on target separation and target size in the display. However, the decision on these parameters cannot be made before taking into account their effect on interaction. Target size and target affect time and accuracy of selection from an interaction point of view. Increased target width leads to increased inter-element distance. An understanding of importance of the two variables relative to each other is therefore necessary to proceed with display design in a formal way.

3. INTERACTION

The action of physical pointing can be performed using different media such as our hand, our head or a hardware device controlling a virtual pointer. As has been found in [18], depending on the resolution of the gesture that is used to point to a sound, the target width has to be adjusted accordingly and different interaction techniques result in different effective angle spans. Based on the results of [18], in a hand based physical pointing task in a spatial audio in the presence of feedback, 15 degrees target width are expected to result in approximately 67% success. Assuming a slope of one for the associated psychometric function, we would need about 22.4° for a selection rate in the vicinity of 100%. It should be noted that no data on the psychometric function of this particular task have been found by the authors and thus this assumption on the slope of the function is purely hypothetical. However, it is used later on in the study and for this reason it is derived here.

Physical gestures as an interaction technique in human computer interaction have not been examined in detail. Such a study is necessary since the motor properties physical gestures are based on, do not work uniformly for all directions relative to one's body. For example, it can be difficult and error prone to point to a direction to one's back when he is facing forward.

Selecting a feedback marked audio display element based on the direction of the sound event either by a real or virtual pointing gesture has many similarities to homing to a visual target, as in Fitt's law experiments. This is due to the fact that participants are informed on the direction of the sound event by spatial audio and they are assured they are on target by audio feedback. However, a different sensory modality is used for event localization and as has been described in the previous section, users have a less precise impression of target location. Results of studies focusing on visual targets can therefore serve well as a starting point that can help to identify parameters that affect this type of interaction. We decide to base our further analysis on the quantities of distance to target and target width since they have been proven to affect virtually all pointing tasks and serve as a well founded starting point for such an investigation. We hypothesize that interaction in a spatial audio display is affected by the prominent variables of target width and distance to target in a manner similar to what has been described by Fitt's law [16]. However, given the novelty of this interaction technique and the absence of any results in the literature, the pointing based 3D audio selection task has to be examined individually to reach conclusions on the properties of this interaction technique.

According to Fitt's law time to select a target is affected by distance to target and target width, in accordance to Equation 1.

$$MT = \alpha + \beta \log_2\left(2A/W\right) \quad (1)$$

In Equation 1 α and β are constants estimated using linear regression, A is the distance to target and W is target width. The log term is called the index of difficulty (ID) and its unit is bits, due to the base of the logarithm being 2. The reciprocal of β is the index of performance (IP) with unit bits/sec. IP has been associated with the rate of information processing for the movement task under investigation and is therefore treated as measure of the efficiency of different interaction techniques. There is evidence that the formulation in Equation 2 proposed by MacKenzie is prevailing since it gives a better fit with observations, and always gives a positive value for the index of difficulty.

$$MT = \alpha + \beta \log_2\left(A/W + 1\right) \quad (2)$$

Although Fitt's law provides a comprehensive formula for the description of selection activities it has been formulated based on empirical observations of people homing to targets using the visual channel. In this sense, there is no guarantee that the law will apply in a different context. Friedlander *et al.* [10], investigated targeting non-visual targets. In each trial of their experiment, participants were asked to move into one out of four directions while counting certain steps indicating ring widths in a bulls-eye menu. Audio and tactile feedback had been tested as a means of marking ring widths to define target distance in the display. According to the results found by the authors, a linear model was more appropriately accounting for movement times in the case of targeting non-visual targets. This implies that

participants followed a different behaviour in their targeting strategy. The authors verified that distance to target and target width indeed affect time to select. The formula that the authors suggest for the approximation of time to target is

$$MT = \alpha + \beta \cdot A/W \quad (3)$$

In Fitt's law type of experiments participants were in direct and continuous contact with the target. In the Friedlander study however, participants were only informed on the direction they should move and the number of concentric circles they had to cross. The fundamental difference between the two tasks is that in Fitt's law studies participants are perceptually aware through their senses of target position, whereas in the Friedlander case participants verify they have reached the target through a cognitive process. This different control option participants used, resulted in a linear relationship between time, distance and target width.

The two aforementioned equations resulted from experiments that differ in modality and interaction technique. For this reason, the experiment that follows is an initial investigation to give an indication on whether any of the laws predicting movement time that have been presented, apply to pointing to an audio target. Furthermore, based on time and accuracy measurements the experiment attempts to answer the display segmentation question that has been already described.

It is case that the relationship being the logarithmic or linear in equations 2,3 does indeed reveal a lot with respect to the control process taking place in the human motor system [20]. However, display design choices such as target size and inter-target separation depends on the relative salience of distance and width rather than on whether the relationship is logarithmic or linear. If the saliency of distance to target is equally important to target width as stands in Equations 2, 3, then any increase in distance followed by an equal increment in target width will not affect time to select the target. If distance has more salience than width in time to select a sound source, then display elements must be placed as close as possible keeping a reasonable target width. On the other hand, if width prevails then it is worth placing the sounds in the display utilizing the whole display area.

4. EXPERIMENTAL TASK AND DESIGN

In order to test which of the hypotheses prevails, we designed two interfaces that are characterized by equal ratios of distance to target and target width. The MINIMAL interface contained four sounds each having target width of 20°. As discussed in Section 3, it could require 22° or more to be on the safe side with respect to selection. However, given the non-existence of psychometric functions for the particular task, we decided to restrict the target width, expecting a slightly higher than one slope for the psychometric function. Sounds were placed every 20° starting from -30° and up to 30°. Sound locations were thus at -30°, -10°, 10°, 30°. The MAXIMAL interface also contained four sounds, each having target width of 45°. Sounds were placed every 45° starting from -67° and up to 67°. The interface was placed on the circumference of a circle in the horizontal plane with 0° in front of the user's nose. Experiment trials were designed to require participants to move between the available position pairs in both interfaces thus resulting in distance arcs of 20°, 40° and 60° for the MINIMAL interface and 45°, 90°, 135° for the MAXIMAL interface. The ratios of distance to target to target width were constant in both displays having values of 1, 2 and 3.

Participants initially had to listen for a target sound, played in isolation from a certain position in space. When finished, the target sound played continuously together with three distracter sounds that were placed in the available display element slots. An angle span was associated with each sound. To select a sound, participants had to point at its location and make a downwards wrist gesture to indicate selection. Participants were instructed to wait until the target sound announcement was finished and then to start moving to the target sound. Participants received audio feedback (the sound of people cheering), emitting from the direction of the target sound, while they were within the target sound's area. An XSENS MT-9B orientation tracker (www.xsens.com) was used to track the orientation of the user's hand and the selection gesture. Participants held the tracker in their palms. To avoid any effects related to timbre, the same sound was used for both distracters and the target sound. This was a short (0.5 sec.) segment of white noise. To improve intelligibility, we introduced a 300ms onset difference between neighbouring sounds. Counting from left to right, this resulted in the second sound starting 300ms later than the first sound, the third 600 ms later and the fourth 900 ms later than the first sound. Sounds repeated after a 500 ms period of silence. All sounds in the display were played in the same level. Sounds were placed according to the MAXIMAL or the MINIMAL specification, in the horizontal plane. In every second trial, the target sound was located in the leftmost position. The location of the target sound for every other task was selected randomly out of the three remaining ones. This resulted in two distance pairs of three distances each with arc length of 45°, 90° and 135° and 20°, 40° and 60° arc length respectively.

There was a short training session prior to testing, during which participants' performance was monitored to make sure that they understood the task. After participants successfully completed four consecutive trials during the training session, the testing started. Participants were tested in the two experimental conditions associated with their groups, one followed by the other in a counterbalanced order.

	Locations	W	A
MINIMAL	-30°,-10°,10°,30°	20°	20°, 40°, 60°
MAXIMAL	-67°, -22°, 22°, 67°	45°	45°, 90°, 135°

Table 1. Experiment Setup and Experimental Conditions. W stands for Target Width and A for distance to target

Sixteen right-handed participants were tested (17 to 27 years, 2 females and 14 males). Participants were paid for their participation. None of the participants reported any hearing deficiencies, however there was no hearing test performed to verify hearing ability.

Time to complete trials, angular deviation from target and movement pattern for each trial were recorded during the experiment. After testing in each experimental condition participants completed a NASA TLX subjective workload assessment. The experiment lasted half an hour. In total 256 measurements were available per level combination.

5. RESULTS

The analysis involves a within-subjects comparison of all sixteen participants. Data were examined with respect to distance

to target and target width ratios (A/W) and the two display segmentation strategies resulting in a 2x3 analysis of variance.

5.1. Time Analysis

Mean times for selection in both display placements are presented in Figure 1. An ANOVA showed a main effect for interface type, with the MAXIMAL interface significantly faster than the MINIMAL ($F(1,255) = 9.687$, $p = 0.002$). A/W ratios significantly affected the results $F(2,510) = 49.149$, $p<0.001$. Pair-wise comparisons using Bonferroni confidence interval adjustments showed all distance to target and target width ratios differ significantly.

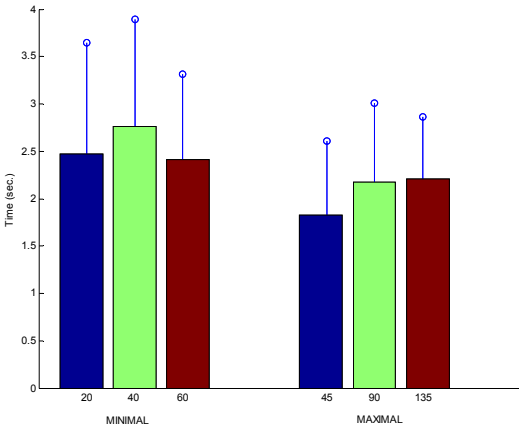

Figure 1. Mean selection times for the MINIMAL and MAXIMAL interfaces.

5.2. Accuracy Analysis

In general participants were successful in selecting within the target area. External feedback proved to be usable and both effective selection angles allowed for efficient selection. The results can be found in Table 2. It is however, the case that the MAXIMAL interface proved to be more accurate $F(1, 15) = 2.96$, $p = 0.011$ and distance to target and target width ratios also influenced the results with the A/W ratio of 1 being more accurate than the other two ($F(2,30) = 9.125$, $p = 0.015$).

	20°/45°	40°/90°	60°/135°
MINIMAL	94.53%	92.19%	87.11%
MAXIMAL	99.61%	94.14%	93.75%

Table 2. Selection success rates for the distances involved in the MINIMAL and MAXIMAL interfaces.

5.3. Additional Observations

Movement trajectories were analyzed and the mean time participants spent in overshooting during all trials and for all sound positions in the experiment were calculated. As can be seen in Figure 2, participants spent more time overshooting in the MINIMAL display arrangement, compared to the MAXIMAL, as would be expected. In particular, the 10° sound location resulted in the highest overshooting time. The MAXIMAL interface resulted in negligible overshooting time for most of the cases, as the targets were large.

In addition, histograms were calculated, for the two display types with respect to the position at which participants indicated selection, see Figure 3. Participants were relatively consistent in their selections and they targeted in a manner that is quite close to the normal distribution. It is also interesting to see that the most frequent selection angles were quite close to the actual sound positions. However, in most of the sound locations, participants were selecting slightly after the target position. A one sample Kolmogorov-Smirnoff test was performed to test for normality of the distributions. In all but the -67° and 30° locations, the histograms proved to follow the normal distribution. The results are summarized in the Table 3.

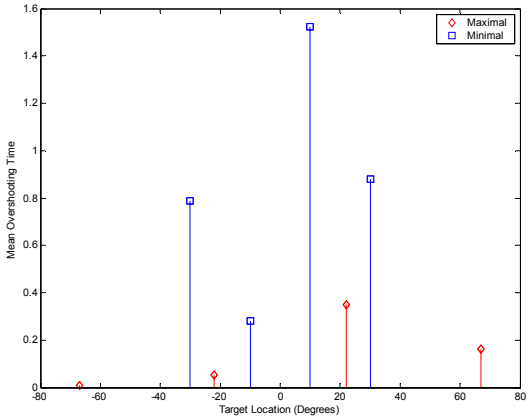

Figure 2. Mean time ratings participants spent in overshooting the target per display type.

Finally, based on the timing results the associated Indexes of Difficulties for the two tasks were calculated through linear regression, according to the model described by Equation 2. In the MAXIMAL case, performing regression on Equation 2 resulted in $a = 1.45$ and $b = 0.4044$, and the regression accounted for 94% of the variance ($r^2 = 0.8887$). Regression was also performed for the linear model in Equation 3. The r^2 statistic was 0.82. We followed the procedure proposed by Friedlander [10] for comparing the goodness of fit of the two models and although Fitt's model accounted for more of the variance, the difference between the goodness of fit of the two models was not found to be significant using the Hotteling's t-test. However, based on r^2 values only it was decided to continue the analysis in the paper based on Fitt's model to be able to compare performance indices with the ones already obtained from other relevant studies in the literature. The Index of Performace for the MAXIMAL case was 2.4728. The data for the MINIMAL case could not be explained satisfactorily by any of the models therefore no further analysis was performed.

\°	-30	-10	10	30	-67	-22	22	67
Z	1.2	.8	1.2	1.5	4.3	1.1	0.8	1.2
S	.12	.53	.13	x	x	.16	.48	.13

Table 3. One sample Kolmogorov-Smirnof Z scores using significance levels determined by the asymptotic distribution.

The standard deviations of selections are presented in Table 4. We observe that the deviation of the selection is higher as the target width increased. Participants adjusted to the larger feedback area and their selections were more spread.

\°	-30	-10	10	30	-67	-22	22	67
	5.2°	4.7°	6.3°	8.9°	21°	11°	15°	16°

Table 4. Standard deviations of selections with respect to target sound position.

6. DISCUSSION

The results of this experiment show that the MAXIMAL interface proved to be significantly faster than the MINIMAL, although participants had to move a longer distance to reach the target. Due to the fact that the ratios of distance to target and target width (A/W) were constant in both display settings, we can conclude that the relative salience of target width was higher than that of distance to target for the two display arrangements. From a design point of view this indicates that it is beneficial to allow the target width to grow when space in the display is available. This result is also justified from a psychoacoustical point of view since increased target width results in increased target separation, a situation that is known to benefit display intelligibility. As is discussed in [25] intelligibility for audio selective and divided attention tasks is improved with higher spatial separation of display elements.

The results from the experiment appear to be in contrast to Fitt's law which would not predict a significant difference between time ratings in interaction in the two displays, given that the ratios of distance to target and target width were the same. It is also the case that the time ratings of the MINIMAL interface arrangement were not consistent with the behaviour predicted by the Fitt's model. An indication of why this might have happened can be found in the time users spent in overshooting. As can be observed in Figure 2, participants often overshoot the target with the MINIMAL interface. In particular, the higher overshooting time was observed in target location 10° which was involved in the distance pair of 40°, thus explaining the fact that this location had the highest mean selection time. There is therefore evidence that target width in the MINIMAL display arrangement might not have been optimal.

Increased overshooting times in the MINIMAL interface can be attributed to movement dynamics. This observation stems from the fact that selection success rate was high, implying that the size of the target was close to optimal in terms of accuracy. However, although the target width associated with the particular sound was enough to allow participants to select within the target, it could not optimally account for the dynamics of the movement. It is quite probable that participants were reaching the feedback area with a relatively high speed that was high enough to result in overshooting. This can also be attributed to a bias in the perception of the target sound position. Although, this is also likely to have occurred in the MAXIMAL interface, the larger target widths were wide enough to allow participants to adjust their movement speed while in the feedback area and avoid overshooting the target. This reduced the homing time and provided more efficient interaction. Target width therefore can serve as a design tool that accounts not only for accurate selection but also for dynamics caused by misperceptions related to the modality the display is presented with. If for example, a user's movement is expected to be performed in a nervous or hasty manner, as for example when mobile, it makes sense to let target widths grow to account for this phenomenon. On examining Figure 2, it is found that most time in overshooting was spent in the sound positions that were in front of the participant, implying that these display areas are more sensitive than the side ones.

Further support to the non-optimality of the MINIMAL display arrangement can be obtained by examining the accuracy ratings in Table 2.

Considering distance averaged accuracy ratings, we see that participants were on average successful in 91% of the trials in the MINIMAL display and 96% of the trials in the MAXIMAL display arrangement. The lower selection rate in the MINIMAL display arrangement provides further evidence that the area of 20° was probably not enough to allow optimal spatial audio target acquisition in the context of this particular experiment. Participants had to concentrate more and for this reason extra time was necessary for them to place their tracking device inside the feedback area. Considering the psychometric function associated with the particular task it is possible that the 20° interval lies prior to the performance ceiling, rather in the area where improvement on selection rates is still possible. This can also be explained from the increment in accuracy that has been observed in the wider interface.

We can therefore observe that as target width increases, it can account for the deficiencies that were presented and were concerned with dynamics and selection rates. If target width is adjusted to a value higher than 20°, time ratings are expected to converge to the case of the MAXIMAL display. In fact, from a certain target width and up the relative salience of distance to target to target width is expected to be the same, given that distance to target remains within reasonable limits.

Distance to target was a factor that affected the timing ratings. In the more normal behaviour observed in the MAXIMAL interface there is an ordering due to distance. From an interaction point of view, increased distance to target will affect the results according to Equation 2. When possible this can be counterbalanced by increasing target width. This way the delay that is predicted by Equation 2 can be avoided.

Participants proved to be reasonably accurate in both display designs with success ratings of approximately 90% or more. As seen from Table 2, the shortest distance is prevailing in terms of accuracy. A small degradation in accuracy is usually found when distance to target is increased. In our particular case this phenomenon can also be explained by considering the gesture involved and the tracking technology used in the experiment. Due to the fact that an orientation tracker was used, participants had to adjust the direction of their palms relative to their arms when making selections. Given that all participants were right-handed we can conclude that this is more conveniently done for targets to the left. When the target is to the right it is hard to fully point to the correct direction without stretching the arm or turning the body. For this reason, when participants are required to move a large distance to reach a target or when performing a selection in a specific direction that is uncomfortable it is advisable to use higher target widths to compensate for accuracy and timing deficiencies.

It is interesting to observe the distributions of selection angles. The majority of selections followed the normal distribution and only sound locations at 30° and -67° were deviating from the normal. In [1], non-directional feedback on target position resulted in wider histograms of selection angles, compared to the non-feedback case, however no test on normality was presented. In this study, 6 out of 8 positions followed a normal distribution. This can be explained by the fact directionality of

both the stimuli and the feedback which was programmed to appear at same position as the target sound.

The MAXIMAL interface was consistent with the Fitt's formulation. This is verified by the regression results which could account for almost 90% of the variance. The performance index for the particular interface was approximately 2.5. This is less compared to the performance indices found by MacKenzie in [17] for pointing using three different devices in a visual target acquisition task. MacKenzie found performance indices of 4.5, 4.9 and 3.3 for pointing to a visual target using a mouse, tablet and trackball devices respectively. It can be argued though that the presence of the distracter sounds and the use of the same sound for elements in the display negatively affected the performance index. In a more user-friendly display the performance index would reasonably be expected to be higher.

Thus, based on the results of this study Fitt's law is a prominent tool that can be applied in spatial audio display design. However, it is noted that in order for Equation 2 to be valid and used to make predictions with respect to interaction in the display, target width and distance have to be constraint in certain ranges of values. With respect to target width this range of values starts from the point where selection performance inhibits a ceiling effect. Knowledge of the psychometric function related to the task under examination is therefore necessary to provide a better understanding of the interaction task under examination and help design a user friendly spatial audio display. In relevant HCI literature, a success rate of 96% is considered sufficient to indicate a ceiling effect. With respect to distance to target, it has to be such that homing to target does not lead the user to positions that are uncomfortable to reach with respect to the interaction technique used. When the target size and separation are within reasonable ranges of values 'normal' performance can be expected which will be varying in a similar manner to the performance described by Fitt's as this has been formulated in visual target acquisition tasks.

Based on the results of this study it is therefore concluded that spatial audio target acquisition can be examined using the methodology that has been developed for visual selection tasks. A common ground for the comparison of interaction techniques irrespective of the modality they are performed with can therefore be established.

The sound stimulus that was used in the experiment was white noise for all display elements. This is beneficiary with respect to sound localization since white noise is a broadband type of stimulus and also inhibits a great deal of spectral variation over time. The above characteristics are considered to assist sound localization. In a spatial audio display used to accomplish human computer interaction white noise does not provide a good solution for element sonification. Other types of stimuli are more appropriate since white noise is not suitable for delivering semantic information. For a display designed to align elements based on azimuth, it is not expected that localization performance will diminish much when other types of sound stimuli are used. In fact no confusions or other deficiencies have been observed for sounds that are constrained with respect to elevation as in the display used in this particular study. In this sense, although localization might be slightly worse depending on the sounds in the display, the overall direction of the stimuli will be recognized. This fact, in combination with feedback marked display elements will result in interaction behaviour similar to the one observed in this study. If elevation was to be used the situation might become problematic since confusions are more likely to happen. The presence of audio feedback however will certainly help alleviate this problem, search time for display elements however might become an issue if confusions are common in the display.

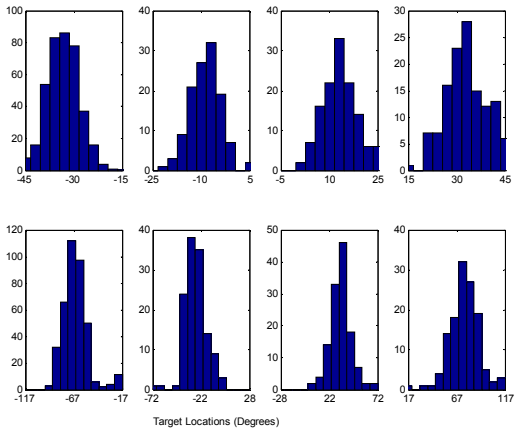

Figure 3. Histograms of selection angles for all targets in the experiment.

7. CONCLUSIONS

This paper presented a study of a spatial audio display design that examined audio target size based on the relative salience of distance to target to target width. The results showed that target width was more important than distance to target in the context of pointing-based gesture interaction with a spatial audio display. Increased target size improved time ratings in the spatial sound selection task and was found to be a useful tool in accounting for misperceptions of sound source positions and direction incurred weaknesses of the motor mechanisms that support physical gestures. The use of a spatially positioned sound as audio feedback resulted in a normal distribution accounting for the participants' selections. The results showed that, given a sufficiently large target size, spatial audio target acquisition in the presence of audio feedback can be modelled using the Fitt's formulation. This enables the direct comparison of interaction techniques for spatial audio target acquisition and modality independent comparisons of interaction techniques. The Index of Performance for gesture based spatial audio target acquisition was found to be less than for virtual pointer visual selection tasks, however time and accuracy ratings support the development of real world applications of spatial audio displays. The results of this paper show that deictic gesture based interaction with a spatial audio display in the presence of audio feedback is a robust and efficient technique.

8. ACKNOWLEDGEMENTS

This study was supported by the Audioclouds project (www.audioclouds.org), EPSRC grant number GR/R98105.

9. REFERENCES

[1] Akamatsu, M., MacKenzie, S. I., and Hasbrouc, T., A Comparison of Tactile, Auditory and Visual Feedback in a Pointing Task using a Mouse-Type device. Ergonomics, 1995. 38: p. 816-827.

[2] Arons, B., A Review of the Cocktail Party Effect. Journal of the American Voice I/O Society, 1992. 12: p. 35-50.

[3] Begault R., D., Wenzel M., E., and Anderson R., M., Direct Comparison of the Impact of Head Tracking, Reverberation and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source. Journal of the Audio Engineering Society, 2001. 49(10): p. 904-916.

[4] Blattner, M. M., Sumikawa, D. A., and Greenberg, R. M., Earcons and Icons: Their Structure and Common Design Principles. Human-Computer Interaction, 1989. 4(1): p. 11-44.

[5] Blauert, J., Spatial Hearing: The psychophysics of human sound localization. 1999: The MIT Press.

[6] Brewster , S., Lumsden , J., Bell, M., Hall, M., and Tasker, S. Multimodal 'Eyes-Free' Interaction Techniques for Wearable Devices. in ACM CHI, 2003. Fort Lauderdale, FL: ACM Press, Addison-Wesley. p. 463-480

[7] Brewster, S. A., The design of sonically-enhanced widgets. Interacting with Computers, 1998. 11(2): p. 211-235.

[8] Bronkhorst W., A., Localization of real and virtual sound sources. The Journal of the Acoustical Society of America, 1995. 98(5): p. 2542-2553.

[9] Cohen , M., Throwing, pitching and catching sound: audio windowing models and modes. Int. J. Man - Machine Studies (1993), 1993. 39: p. 269 - 304.

[10] Friedlander, N., Schlueter, K., and Mantei, M. Bulls eye! When Fitt's Law doesn't fit. in ACM CHI, 1998. Los Angeles, CA: ACM Press Addison-Wesley. p. 257-264

[11] Gaver , W. W., The SonicFinder: An Interface that uses Auditory Icons. Human-Computer Interaction, 1989. 4: p. 67-94.

[12] Goose, S. and Moller, C. A 3D Audio Only Interactive Web Browser: Using Spatialization to Convey Hypermedia Document Structure. in 7th ACM international conference on Multimedia, 1999. Orlando, Florida, United States: ACM Press. p. 363 - 371

[13] Kobayashi, M. and Schmandt, C. Dynamic Soundscape: mapping time to space for audio browsing. in SIGCHI conference on Human factors in computing systems, 1997. Atlanta, Georgioa, United States: ACM Press. p. 194 - 201

[14] Loomis, J. M., Hebert, C., and Cocinelli, J. G., Active Localization of Virtual Sounds. Journal of the Acoustical Society of America, 1990. 88(4): p. 1757-1764.

[15] Ludwig, L., Pincever, N., and Cohen, M. Extending the notion of a window system to audio. in IEEE Computer, 1990. p. 66-72

[16] MacKenzie, S. and Buxton, W. Extending Fitt's Law to Two-Dimensional Tasks. in Conference on Human Factors and Computing Systems., 1992. Monterey, California, United States. p. 219-226

[17] MacKenzie, S. I., Sellen, A., and Buxton, W. A Comparison of Input Devices in Elemental Pointing and Dragging Tasks. in Conference on Human Factors in Computing Systems, 1991. New Orleans, Louisiana, United States: ACM Press. p.

[18] Marentakis, G. and Brewster, S., A. A study on gesture interaction with a 3D Audio Display. in Mobile HCI, 2004. Glasgow, UK. p.

[19] Raman, T. V., Auditory Interfaces: Towards the speaking computer. 1997: Kluwer Academic Publishers.

[20] Richard J. Jagacinski, J. M. F., Control Theory for Humans. 2003, London: Lawrence Erlbaum Associates, Publishers.

[21] Savidis, A., Stephanidis, C., Korte, A., Rispien, K., and Fellbaum, C. A generic direct-manipulation 3D auditory environment for hierarchical navigation in non-visual interaction. in ACM ASSETS '96, 1996. Vancouver, Canada, 1996: ACM Press. p. 117-123

[22] Sawhney, N. and Schmandt, C., Nomadic Radio: Speech and Audio Interaction for Contextual Messaging in Nomadic Environments. ACM Transactions on Computer-Human Interaction, 2000. 7(3): p. 353-383.

[23] Schmandt, C. Audio hallway: a virtual acoustic environment for browsing. in 11th annual ACM symposium on User interface software and technology, 1998. San Francisco, California, United States: ACM Press. p. 163 - 170

[24] Schmandt , C., Ackerman, M. S., and Hindus , D., Augmenting a Window System with Speech Input. IEEE Computer, 1990.

[25] Shinn-Cunningham B. and Antje, I. Selective and Divided Attention: Extracting Information from Simultaneous Sound Sources. in International Conference on Auditory Display, 2004. Sydney, Australia. p.

[26] Wenzel M., E. Effect of Increasing System Latency on Localization of Virtual Sounds. in Audio Engineering 16th International Conference on Spatial Sound Reproduction, 1999. Rovaniemy, Finland: New York: Audio Engineering Society. p. 42-50

[27] Wenzel M., E., Marianne, A., Kistler, J. D., and Wightman, L. F., Localization using non-individualized head-related transfer function. Journal of the Acoustical Society of America, 1993. 94(1): p. 111-123.